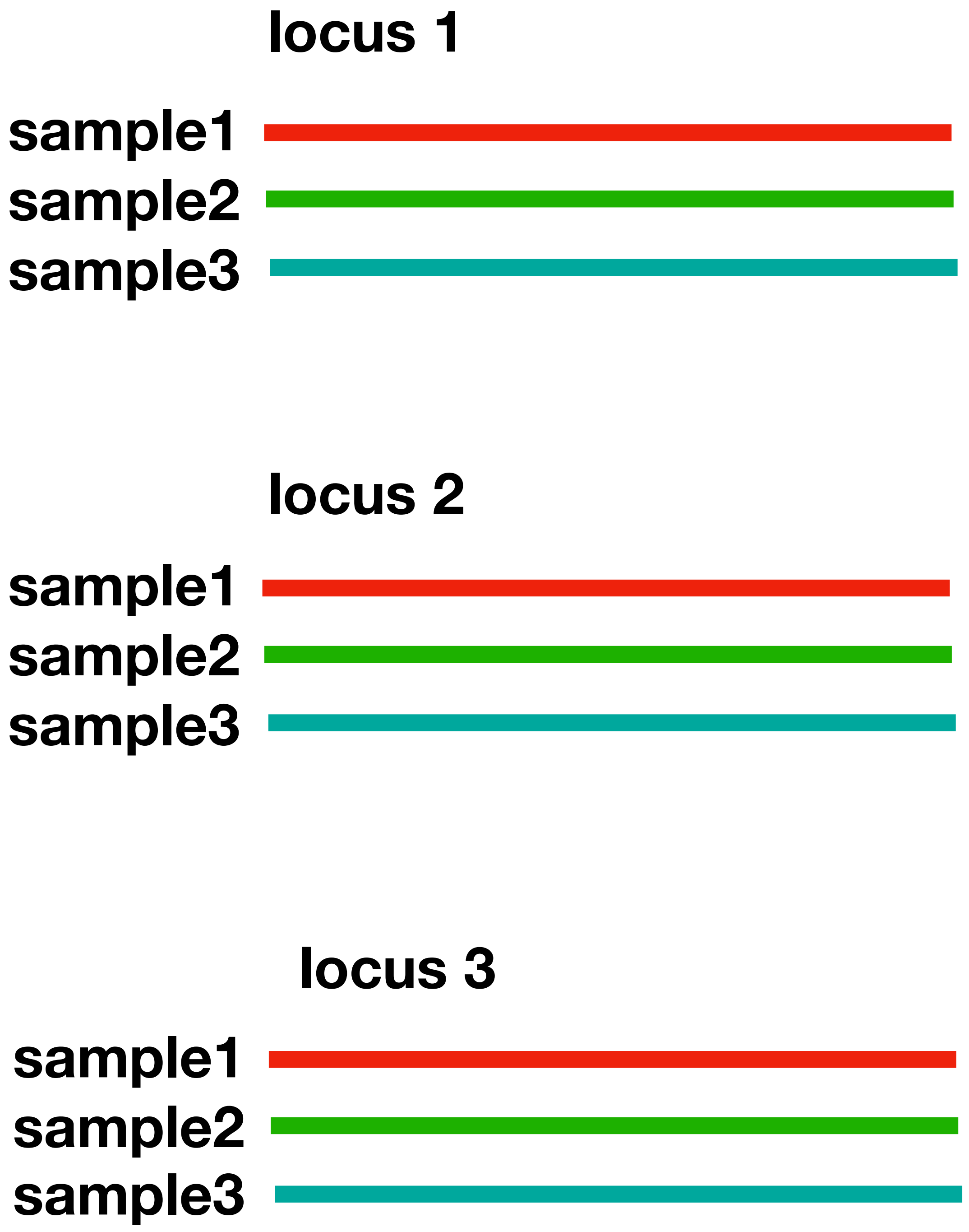
# Post-processing of enrichment data

Reporter: Hao Yuan

# Output of assembling pipeline



locus 1

sample1

sample2

sample3

locus 2

sample1

sample2

sample3

locus 3

sample1

sample2

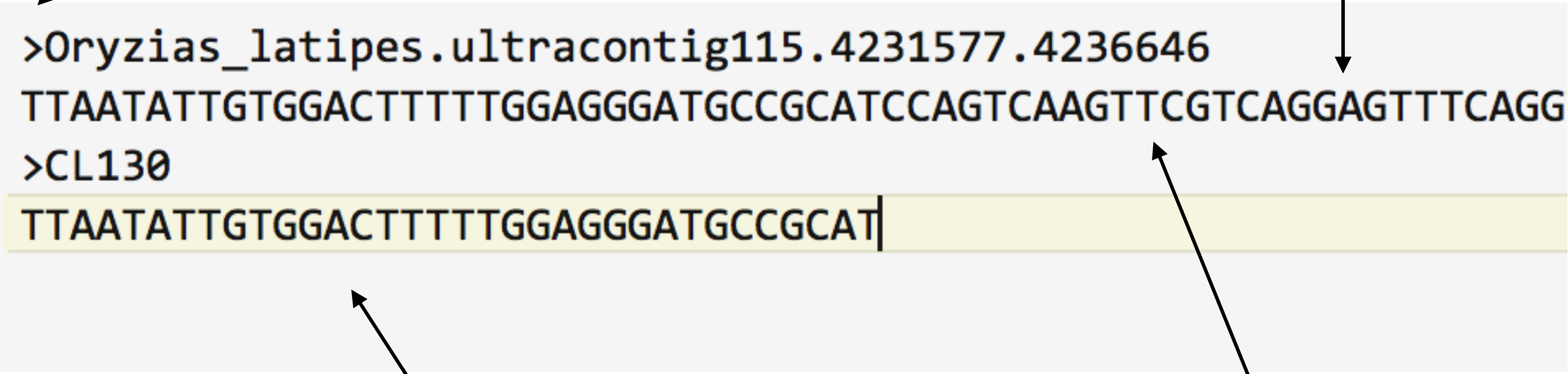
sample3

Sequences targeting the same loci

# Output until here looks like this

A “>” before the sequence name

After name is sequence



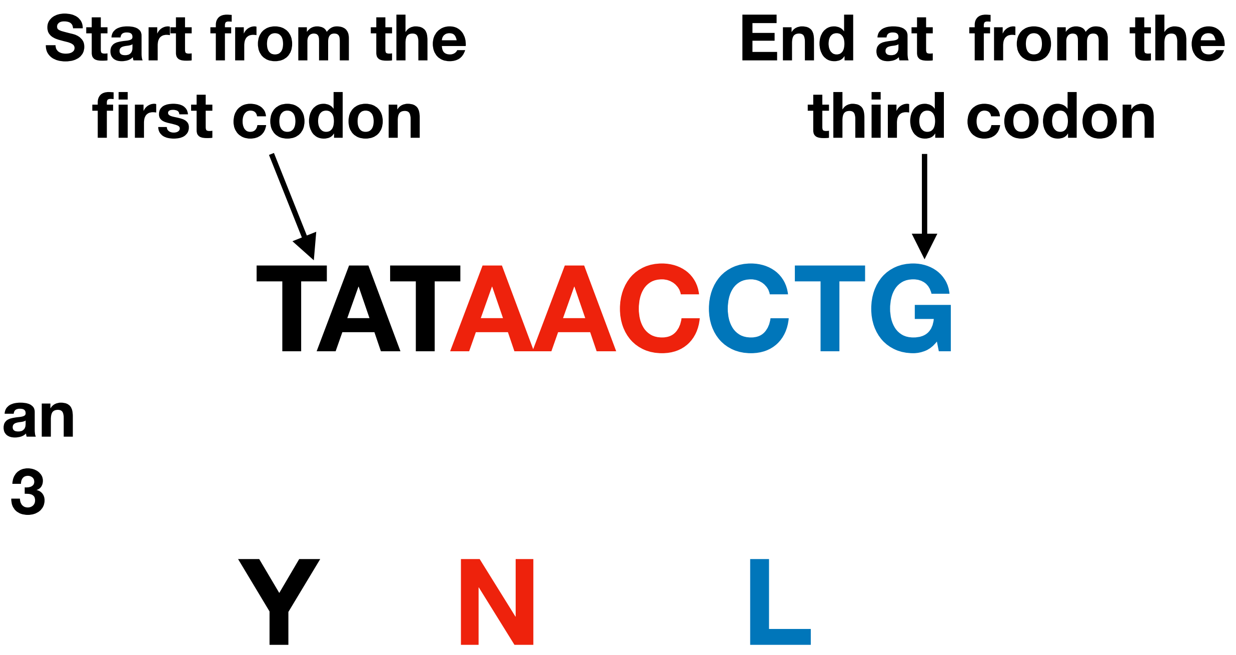
```
>Oryzias_latipes.ultracontig115.4231577.4236646
TTAATATTGTGGACTTTTTGGAGGGATGCCGCATCCAGTCAAGTTCGTCAGGAGTTTCAGG
>CL130
TTAATATTGTGGACTTTTTGGAGGGATGCCGCAT
```

First one is the sequence of reference

Following is the sequence of enriched sample

This called fasta format. File suffix is “fa”, “fas” or “fasta”

## Assembled contigs are coding sequences



Start from the first codon

End at  from the third codon

TATAACCTG

Length of nucleotide can be exactly divided by 3

Y N L

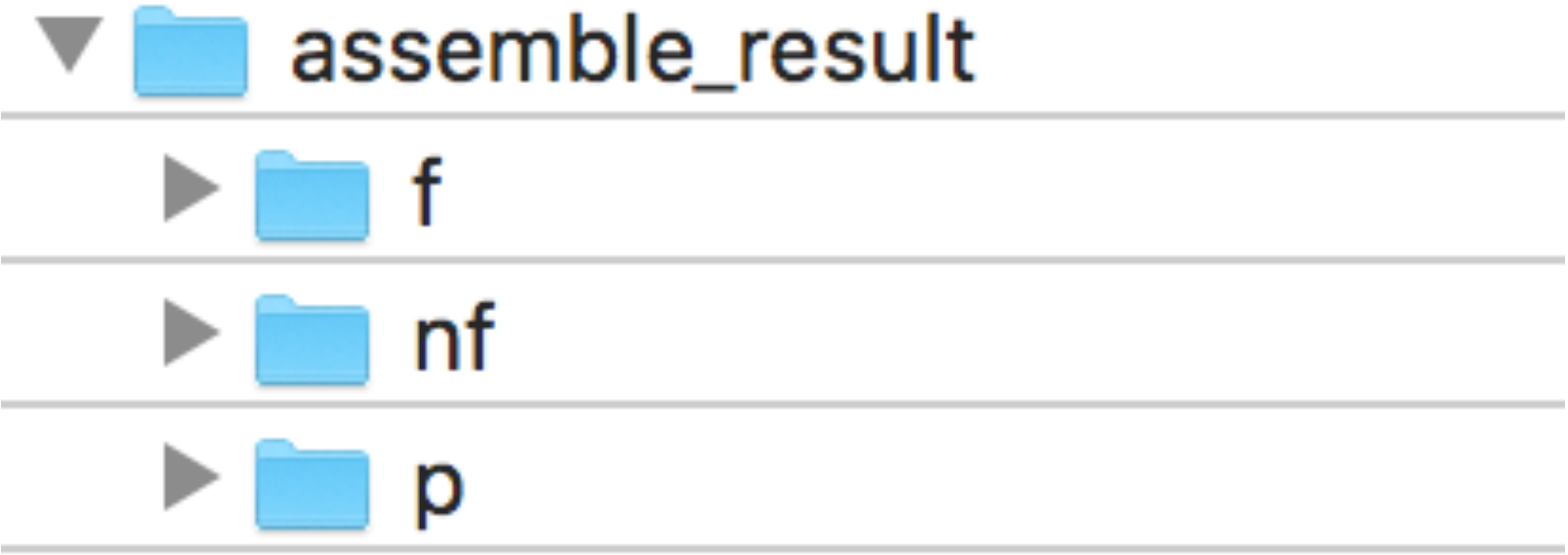
No stop codon in amino acid sequences

Output will be placed under "assemble_result" including 3 folders:

1) nf: folder containing full coding nucleotide sequences

2) f: folder containing coding sequences with flankings

3) p: folder containing amino acid sequences



# Further processing

Data manipulation

Multiple sequences alignment

Filter poorly aligned regions

Filter for other purpose

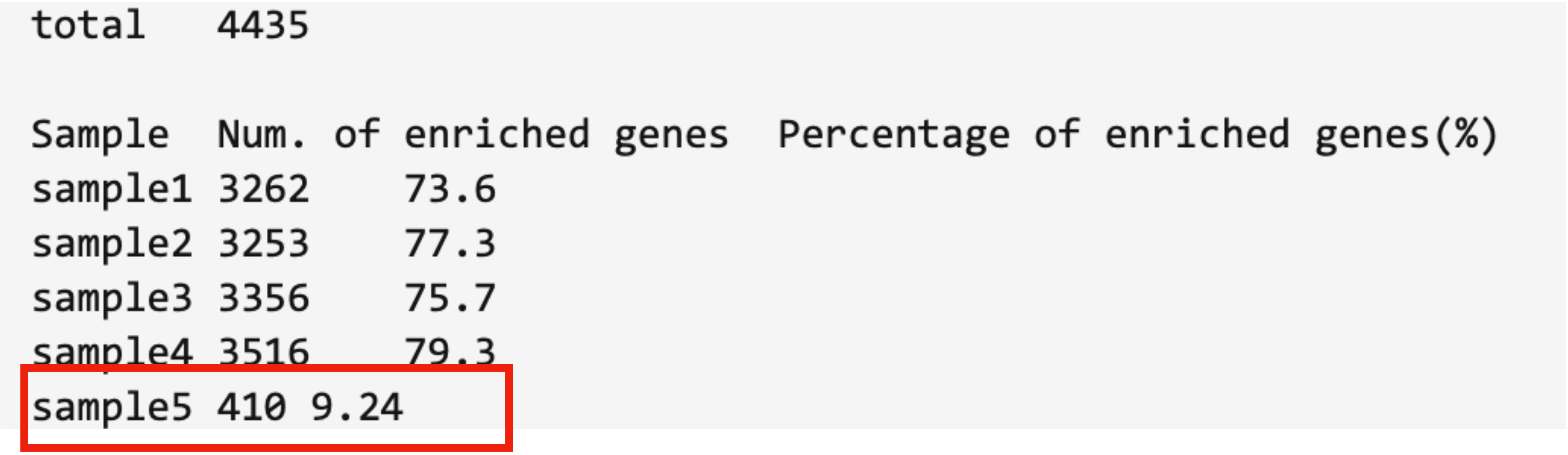
Detect cross-sample contamination

Summary statistics

# Data manipulation

## Remove poorly enriched taxa

enriched_gene.txt



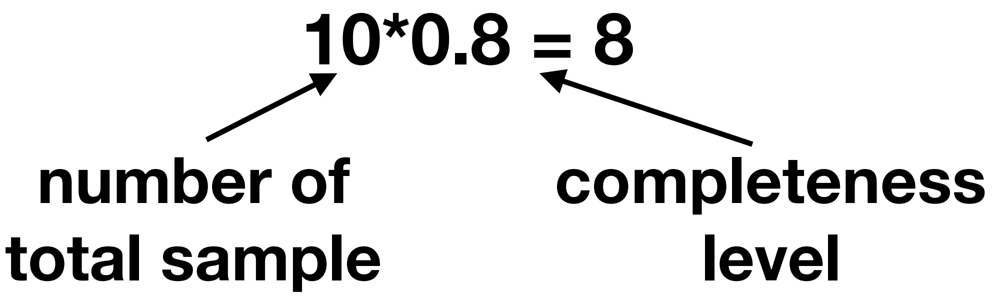
```
total   4435

Sample  Num. of enriched genes  Percentage of enriched genes(%)
sample1 3262    73.6
sample2 3253    77.3
sample3 3356    75.7
sample4 3516    79.3
sample5 410 9.24
```

"—deselected_taxa" option of pick_taxa.pl

## Get loci with certain data completeness level

"—min_seq" option of pick_taxa.pl



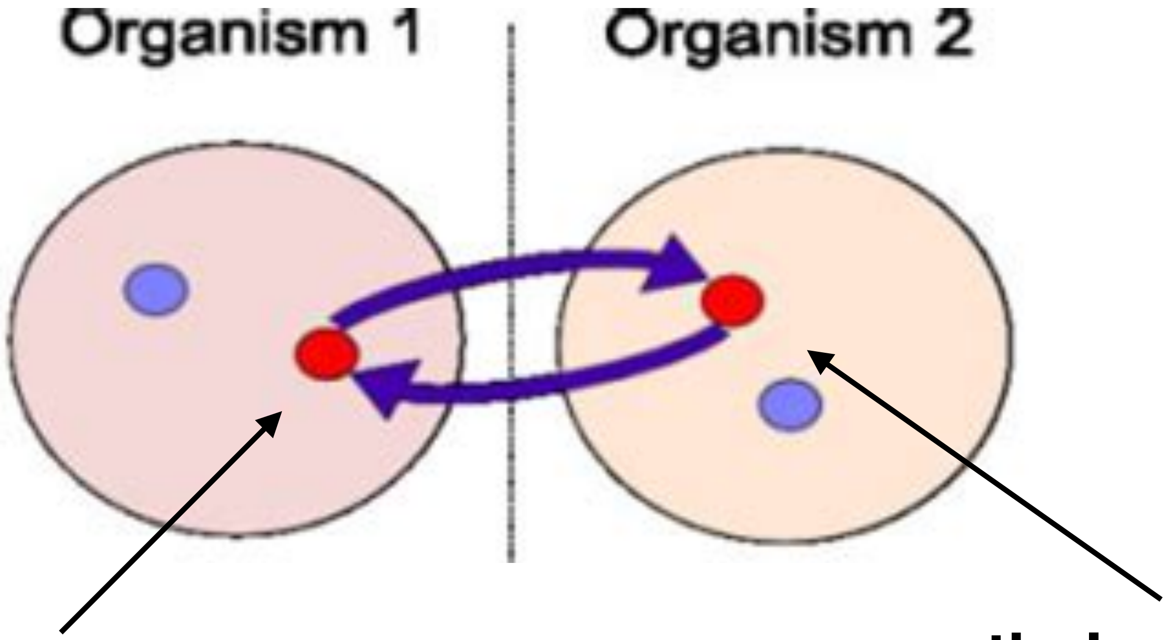
10*0.8 = 8

number of
total sample

completeness
level

# Data manipulation

## Extract loci from existing genomes



one of locus
in reference

orthologous
locus in targeted
genome

use merge_loci.pl to add extracted
sequences to enriched dataset

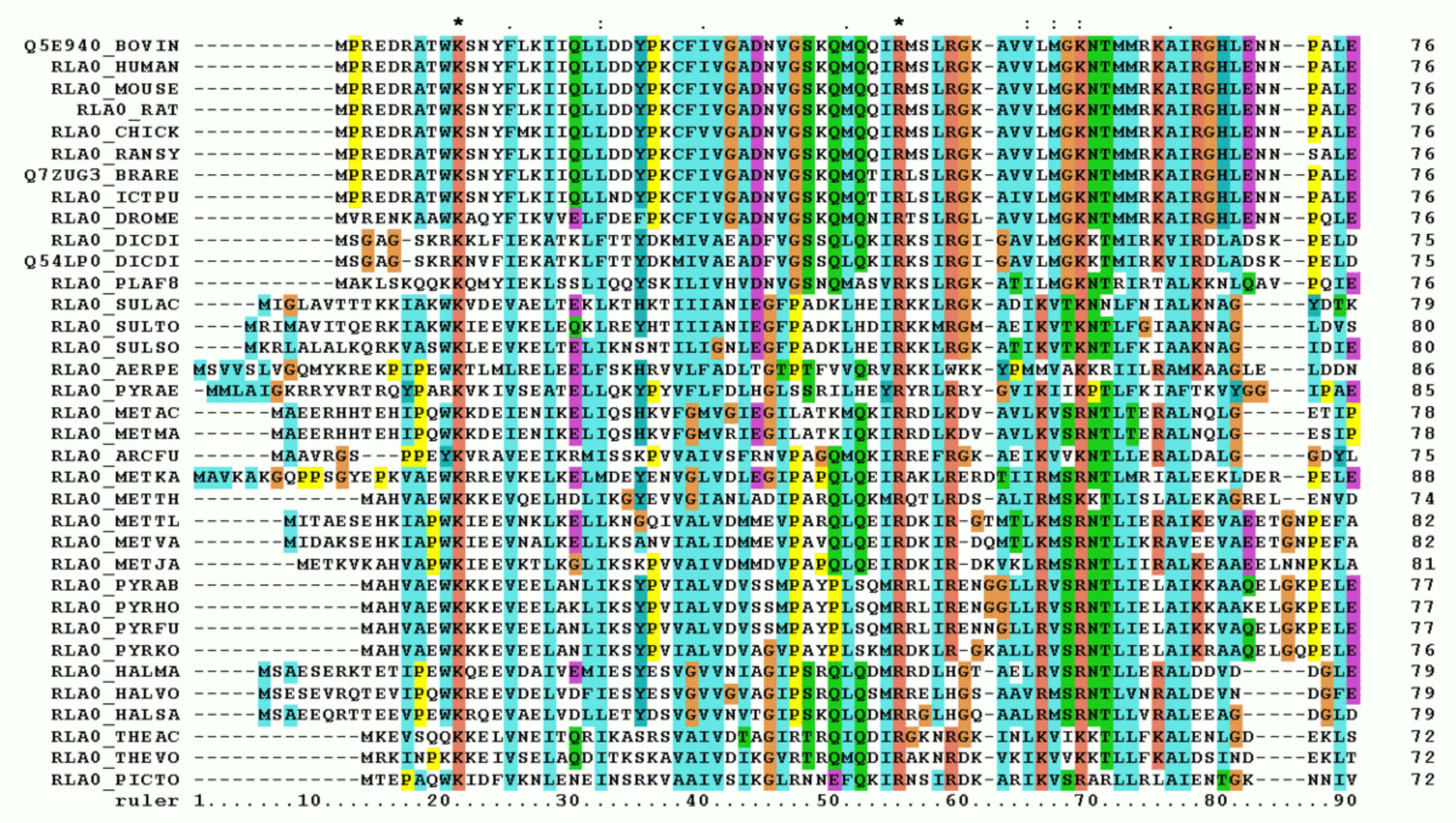
get_orthologues.pl

merge_loci.pl

# Multiple sequences alignment

specify "—non_codon_aln" option of mafft_aln.pl if align sequences with flanks

Reads → Contigs



```
                                 *   .      :       .          .   *         : : :      .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE     76
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE     76
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE     76
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE     76
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE     76
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE     76
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE     76
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE     76
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE     76
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD     75
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD     75
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE     76
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK     79
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS     80
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE     80
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN     86
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE     85
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP     78
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP     78
  RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL     75
  RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE     88
  RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD     74
  RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA     82
  RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA     82
  RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA     81
  RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE     77
  RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE     77
  RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE     77
  RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE     76
  RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD-----DGLE     79
  RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN-----DGFE     79
  RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----DGLD     79
  RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS     72
  RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT     72
  RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV     72
       ruler 1.......10........20........30........40........50........60........70........80........90
```

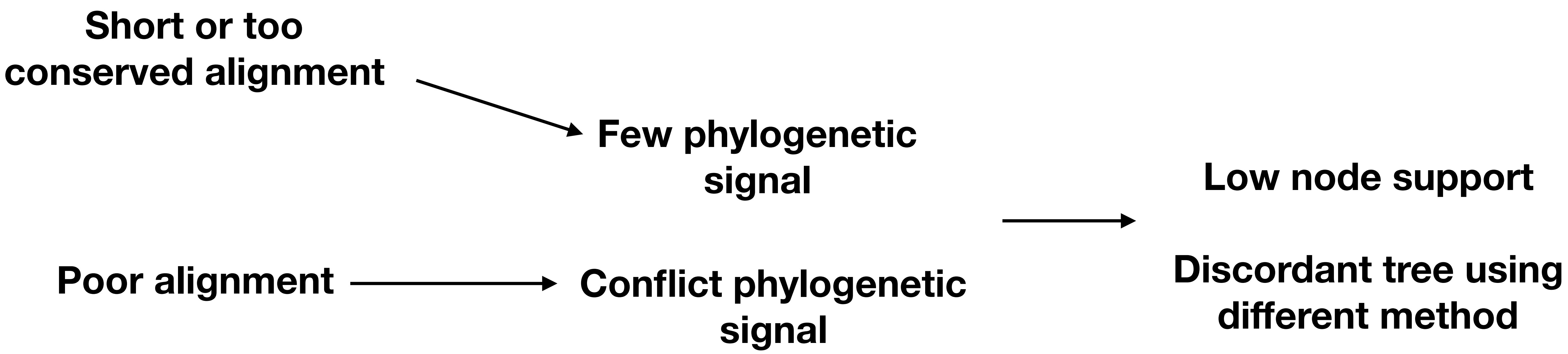
→ Tree reconstruction

mafft_aln.pl

Multiple sequence alignment (MSA)

Align AA/DNA with common ancestry

# Why need to filter resulting alignment



Short or too conserved alignment

Few phylogenetic signal

Poor alignment

Conflict phylogenetic signal

Low node support

Discordant tree using different method

Mis-alignment

Missing data

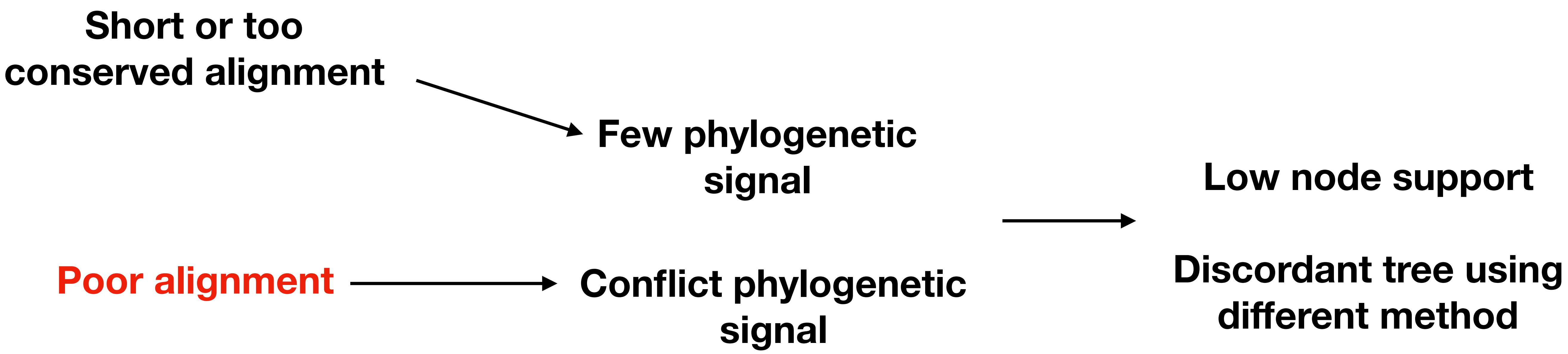
Paralogs

Saturated mutation

```
ATGCGTACGTT
ATGCGTATT--

ATGCGTACGTT
ATGCGTA--TT
```

# Why need to filter resulting alignment



Short or too conserved alignment

Few phylogenetic signal

Poor alignment

Conflict phylogenetic signal

Low node support

Discordant tree using different method

Mis-alignment

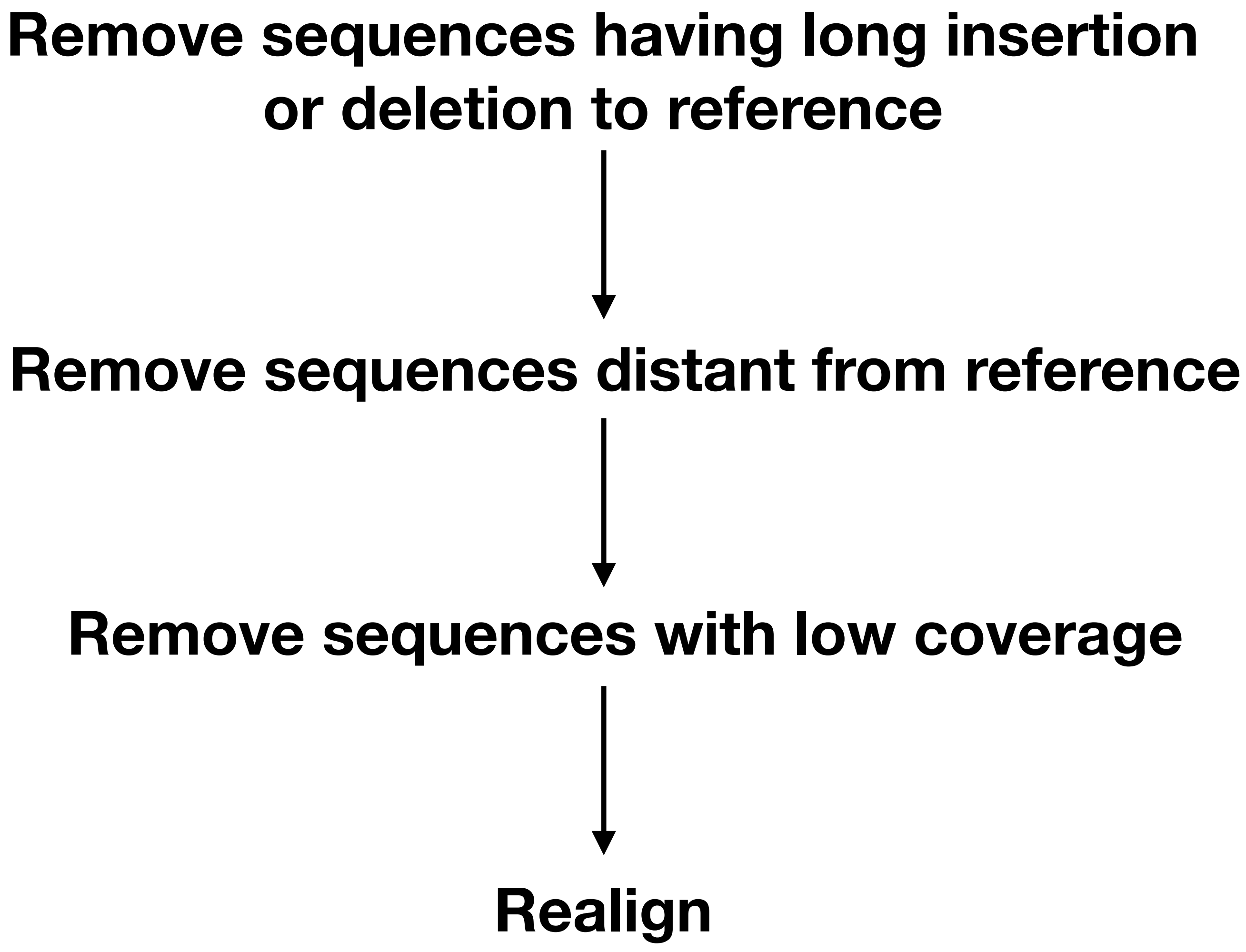
Missing data

Paralogs

Saturated mutation

# Filter poorly aligned regions

## Poorly aligned coding regions



Remove sequences having long insertion or deletion to reference

Remove sequences distant from reference

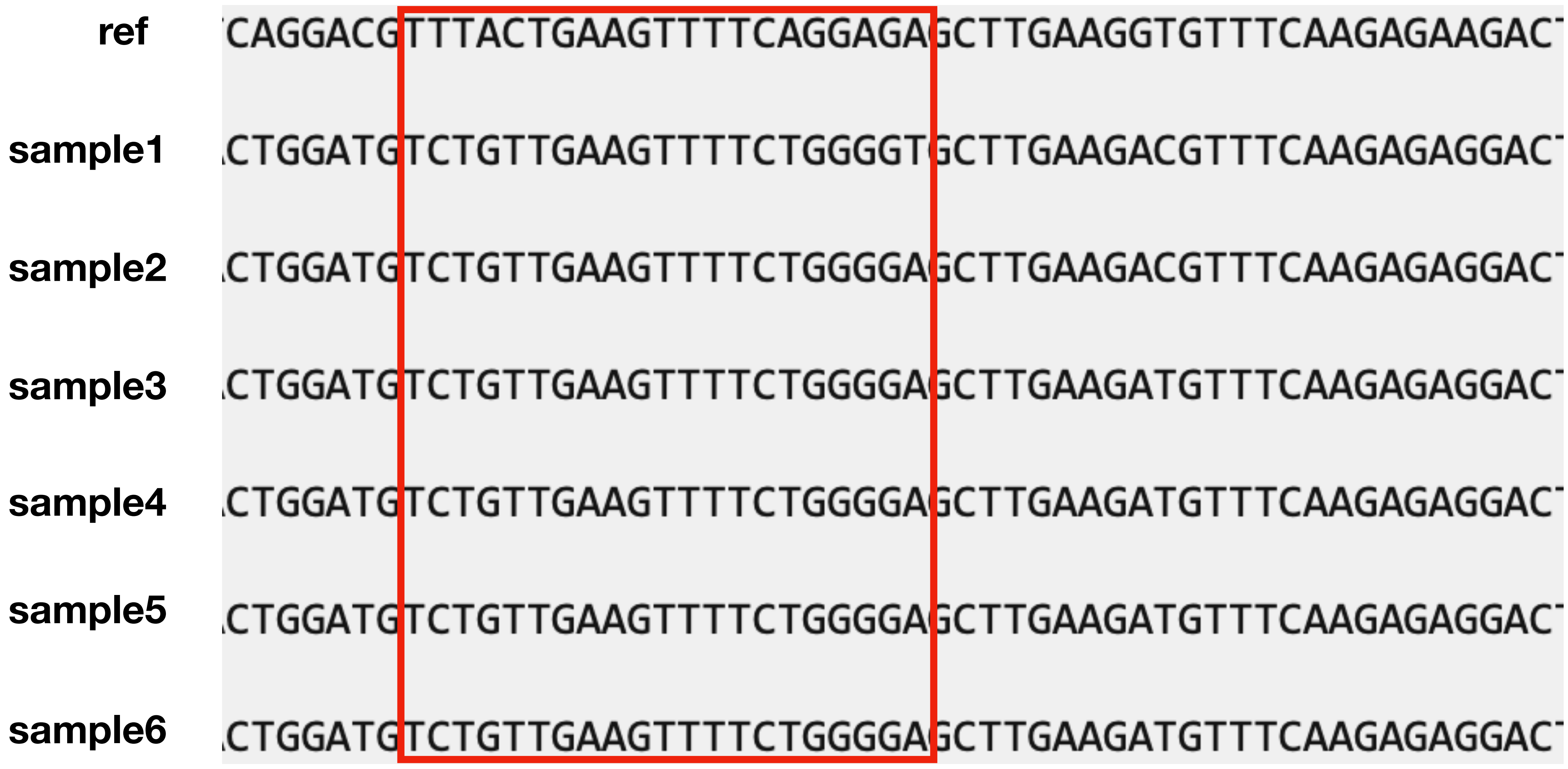
Remove sequences with low coverage

Realign

filter.pl

# Filter poorly aligned regions

## Sequences distant from reference

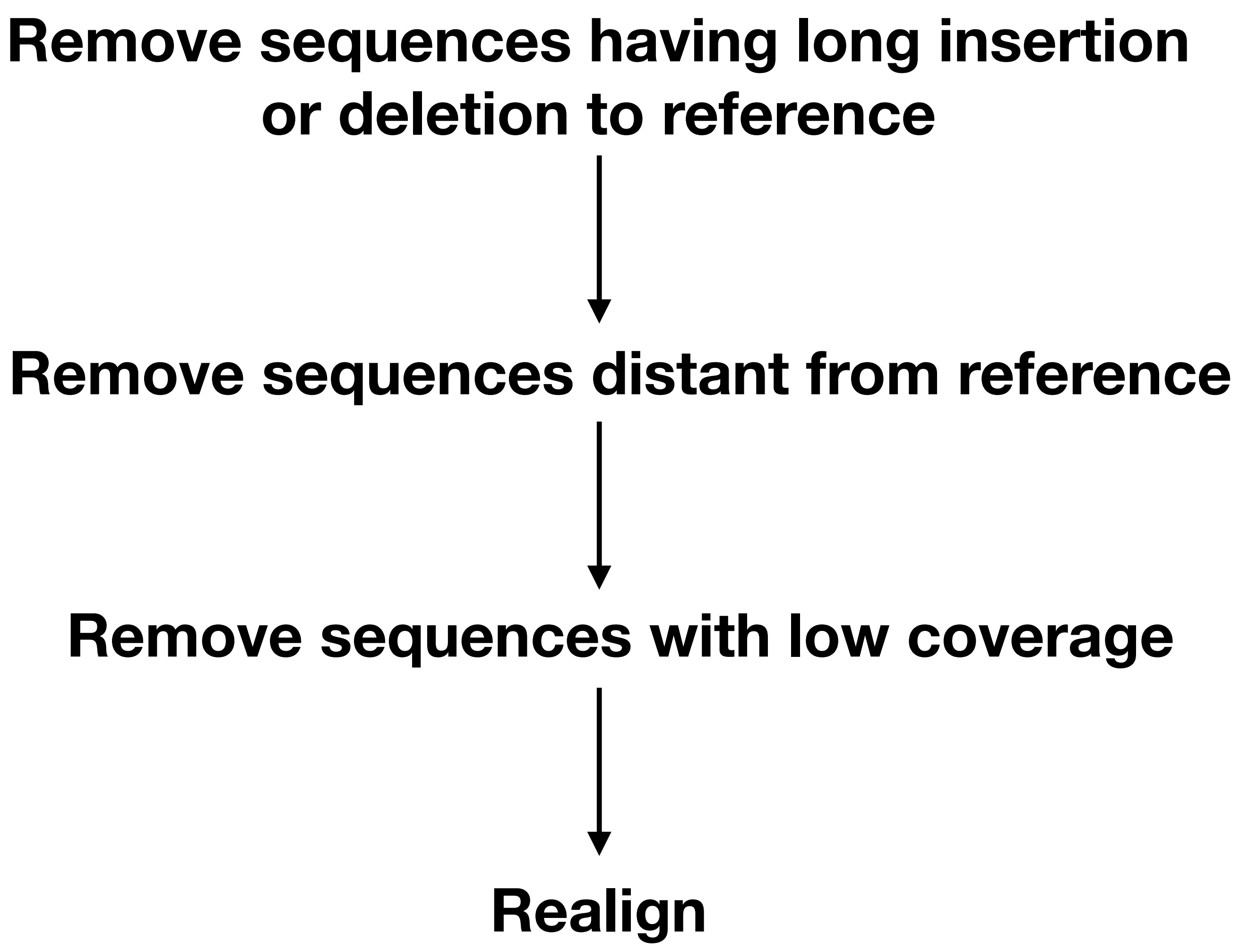


```
ref      CAGGACGTTTACTGAAGTTTTCAGGAGAGCTTGAAGGTGTTTCAAGAGAAGACT
sample1  CTGGATGTCTGTTGAAGTTTTCTGGGGTGCTTGAAGACGTTTCAAGAGAGGACT
sample2  CTGGATGTCTGTTGAAGTTTTCTGGGGAGCTTGAAGACGTTTCAAGAGAGGACT
sample3  CTGGATGTCTGTTGAAGTTTTCTGGGGAGCTTGAAGATGTTTCAAGAGAGGACT
sample4  CTGGATGTCTGTTGAAGTTTTCTGGGGAGCTTGAAGATGTTTCAAGAGAGGACT
sample5  CTGGATGTCTGTTGAAGTTTTCTGGGGAGCTTGAAGATGTTTCAAGAGAGGACT
sample6  CTGGATGTCTGTTGAAGTTTTCTGGGGAGCTTGAAGATGTTTCAAGAGAGGACT
```

**50 bp, 25 bp per step**

# Filter poorly aligned regions

## Poorly aligned coding regions



Remove sequences having long insertion or deletion to reference

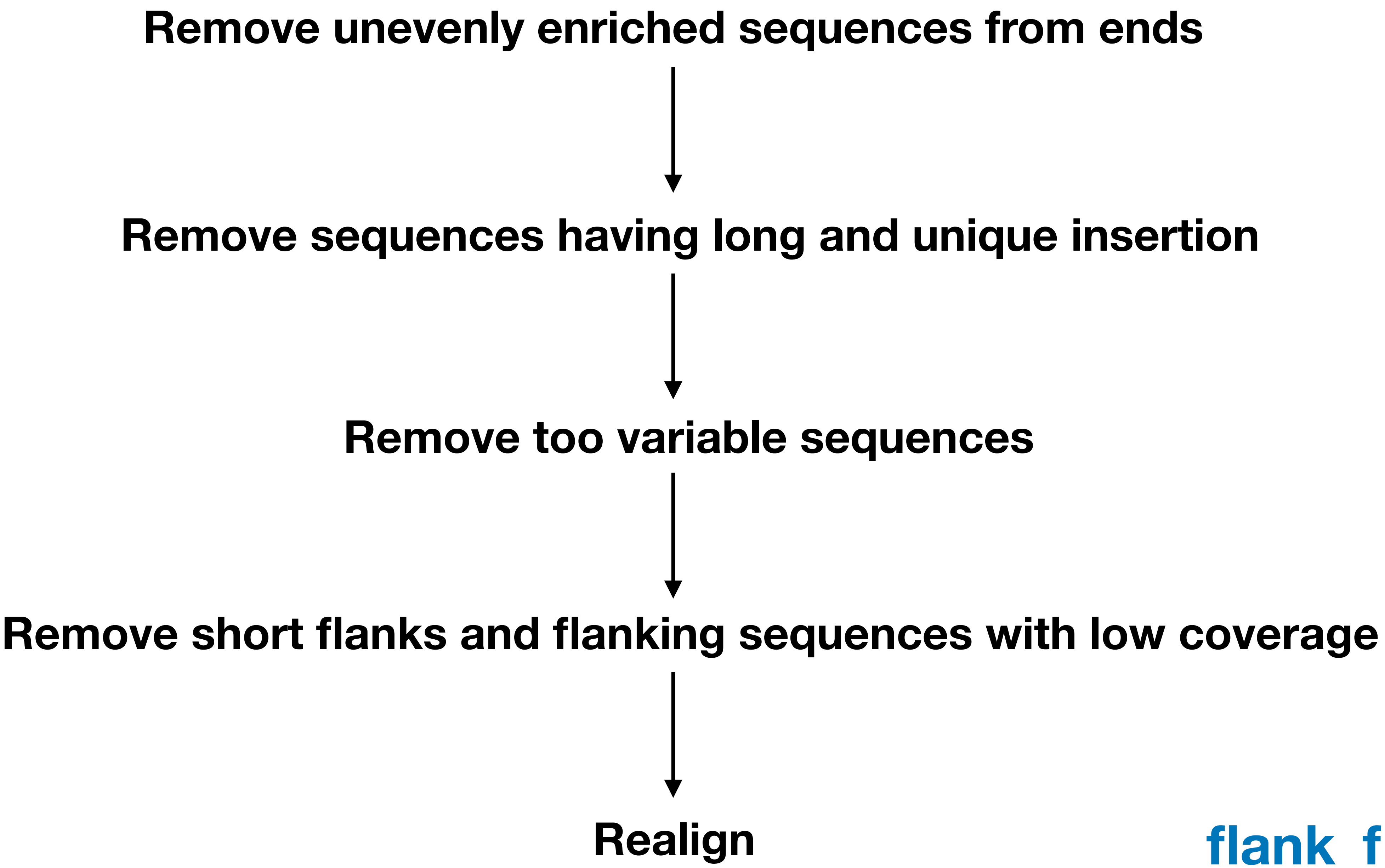
Remove sequences distant from reference

Remove sequences with low coverage

Realign

# Filter poorly aligned regions

## Poorly aligned flanking regions



Remove unevenly enriched sequences from ends

Remove sequences having long and unique insertion

Remove too variable sequences

Remove short flanks and flanking sequences with low coverage

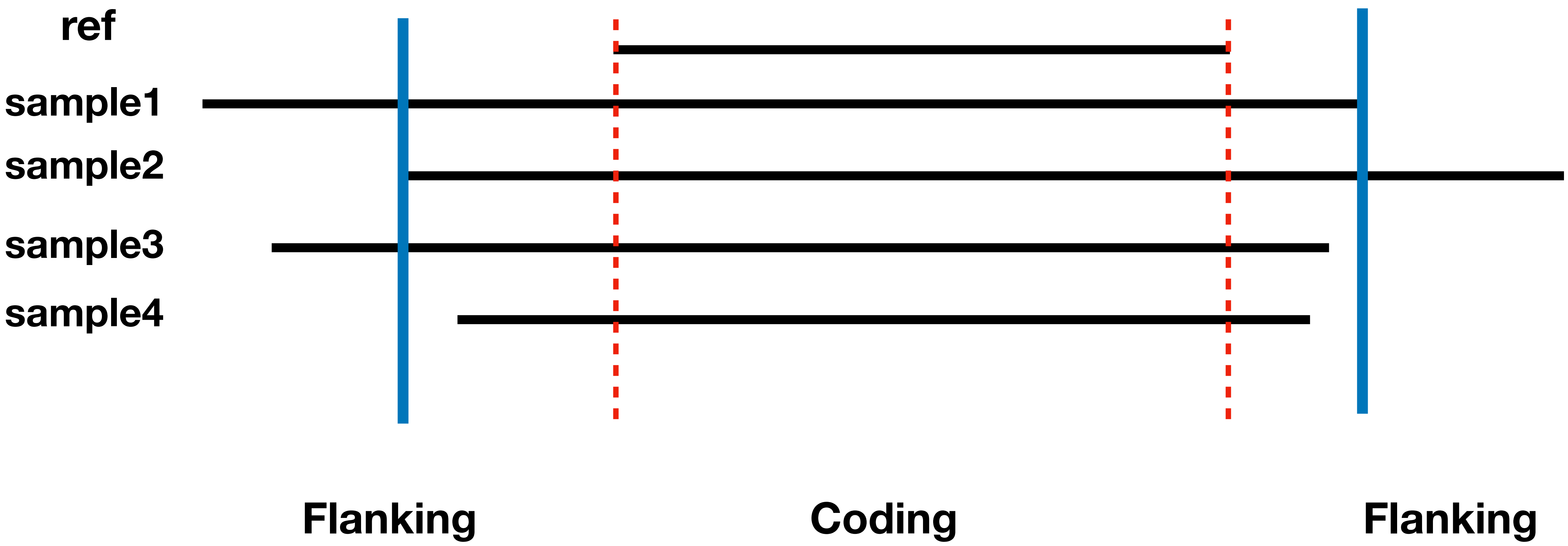
Realign

flank_filter.pl

# Filter poorly aligned regions

## Remove unevenly enriched sequences from ends

consecutive 5 columns
with 50% of residues



ref

sample1

sample2

sample3

sample4

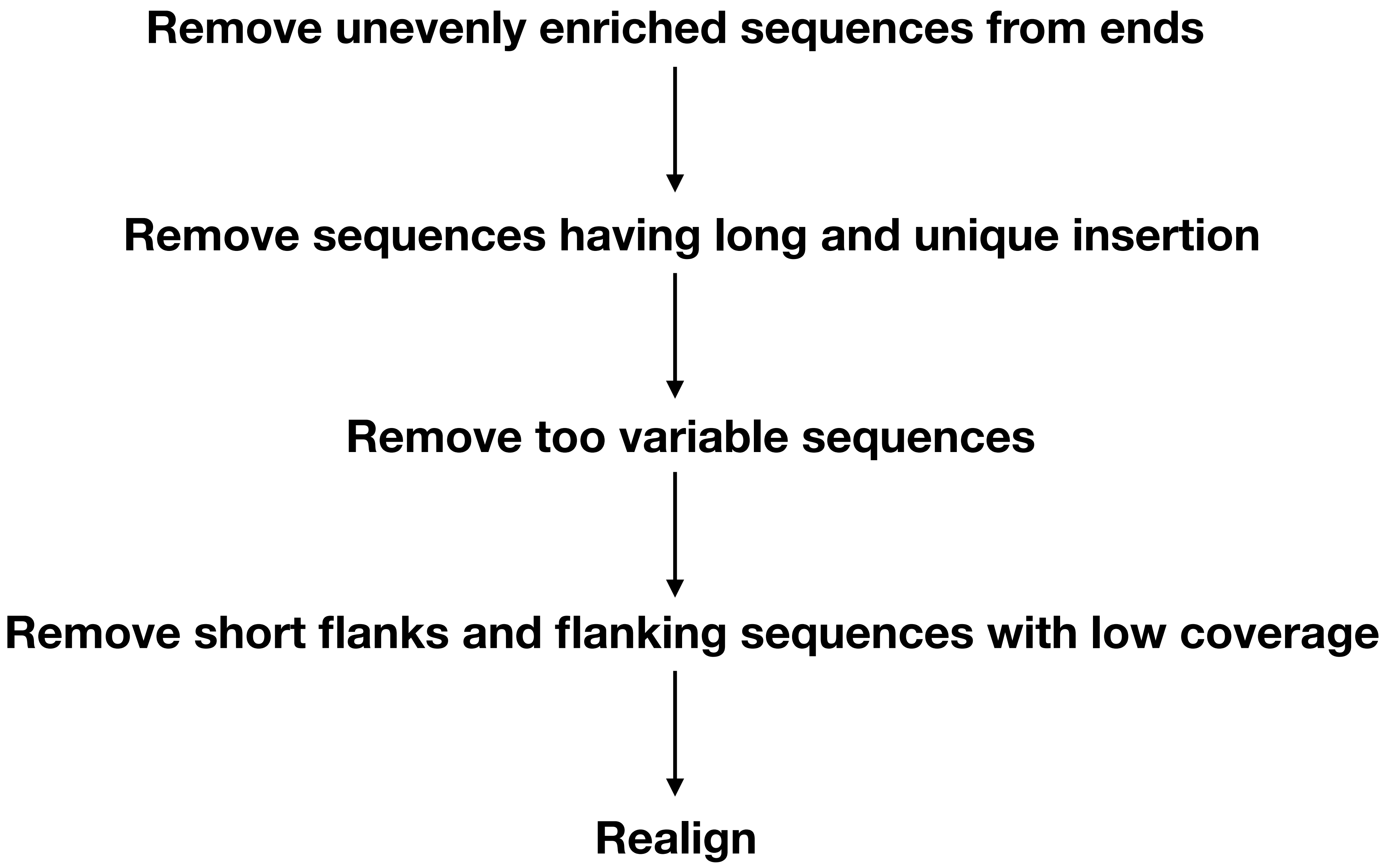
Flanking

Coding

Flanking

# Filter poorly aligned regions

## Poorly aligned flanking regions



Remove unevenly enriched sequences from ends

Remove sequences having long and unique insertion

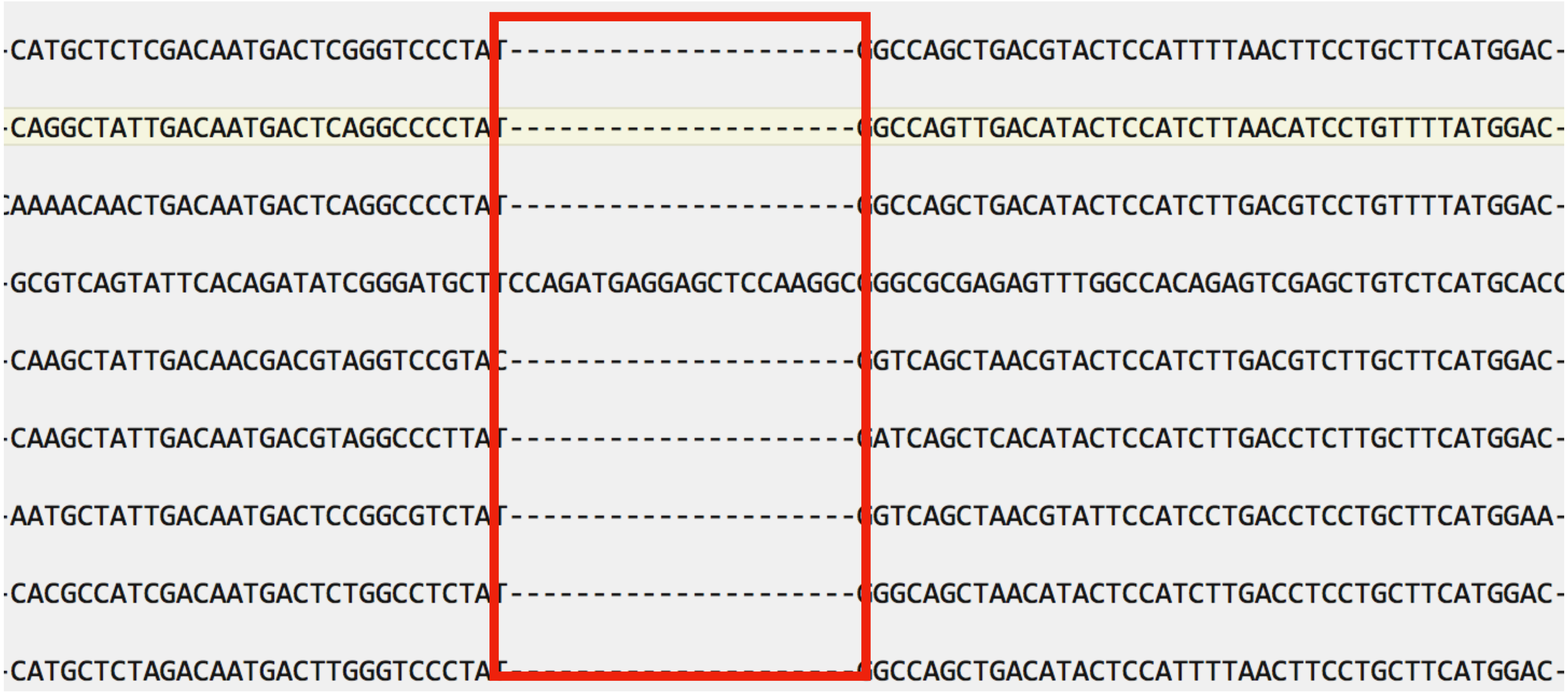
Remove too variable sequences

Remove short flanks and flanking sequences with low coverage

Realign

# Filter poorly aligned regions

## Remove sequences having long and unique insertion



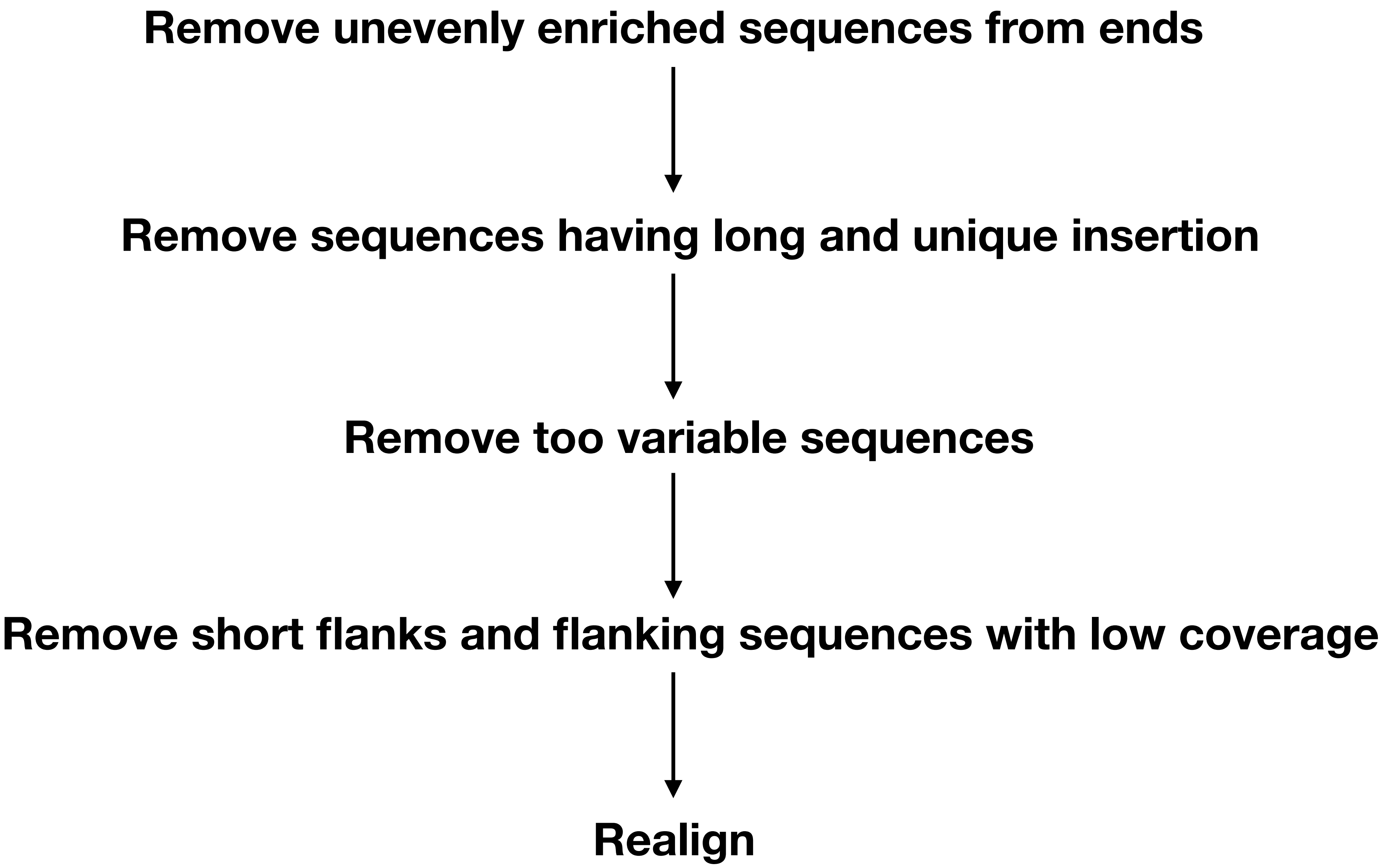
```
-CATGCTCTCGACAATGACTCGGGTCCCTAT---------------------GGCCAGCTGACGTACTCCATTTTAACTTCCTGCTTCATGGAC-
-CAGGCTATTGACAATGACTCAGGCCCCTAT---------------------GGCCAGTTGACATACTCCATCTTAACATCCTGTTTTATGGAC-
CAAAACAACTGACAATGACTCAGGCCCCTAT---------------------GGCCAGCTGACATACTCCATCTTGACGTCCTGTTTTATGGAC-
-GCGTCAGTATTCACAGATATCGGGATGCTTCCAGATGAGGAGCTCCAAGGCGGGCGCGAGAGTTTGGCCACAGAGTCGAGCTGTCTCATGCACC
-CAAGCTATTGACAACGACGTAGGTCCGTAC---------------------GGTCAGCTAACGTACTCCATCTTGACGTCTTGCTTCATGGAC-
-CAAGCTATTGACAATGACGTAGGCCCTTAT---------------------GATCAGCTCACATACTCCATCTTGACCTCTTGCTTCATGGAC-
-AATGCTATTGACAATGACTCCGGCGTCTAT---------------------GGTCAGCTAACGTATTCCATCCTGACCTCCTGCTTCATGGAA-
-CACGCCATCGACAATGACTCTGGCCTCTAT---------------------GGGCAGCTAACATACTCCATCTTGACCTCCTGCTTCATGGAC-
-CATGCTCTAGACAATGACTTGGGTCCCTAT---------------------GGCCAGCTGACATACTCCATTTTAACTTCCTGCTTCATGGAC-
```

chimeric assembly

>= 10bp

# Filter poorly aligned regions

## Poorly aligned flanking regions



Remove unevenly enriched sequences from ends

Remove sequences having long and unique insertion

Remove too variable sequences

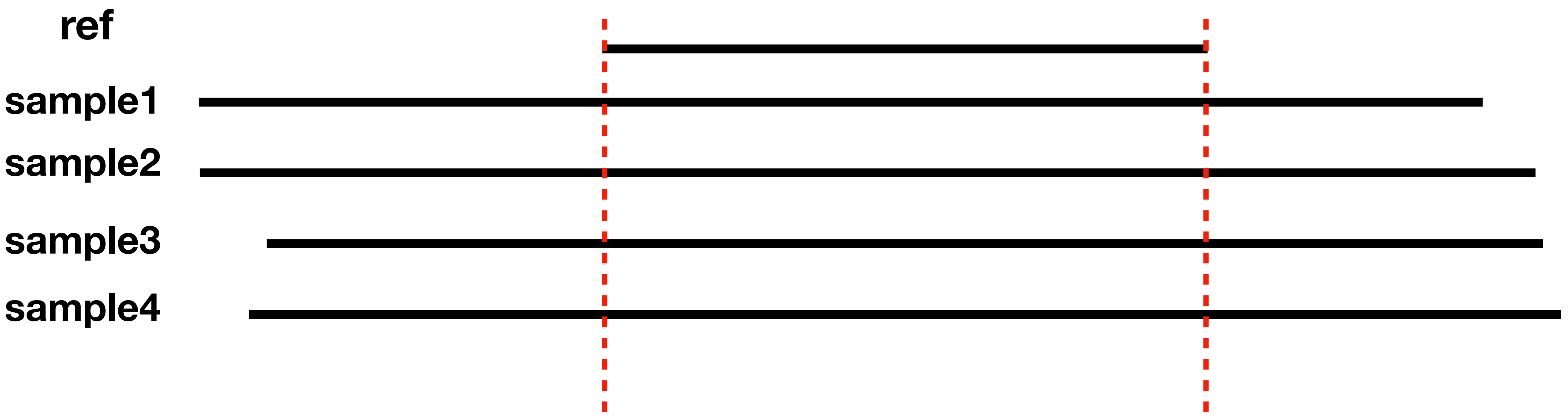
Remove short flanks and flanking sequences with low coverage

Realign

# Filter poorly aligned regions

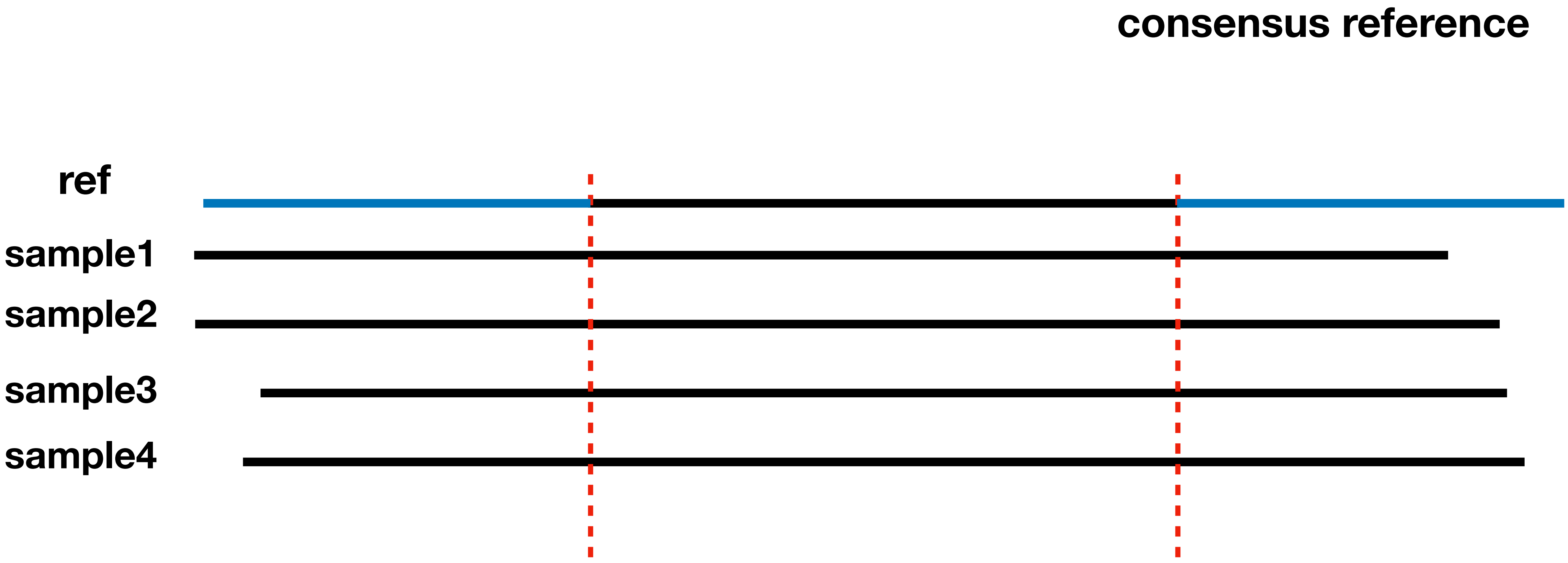
Remove too variable sequences

No reference in flanks



# Filter poorly aligned regions

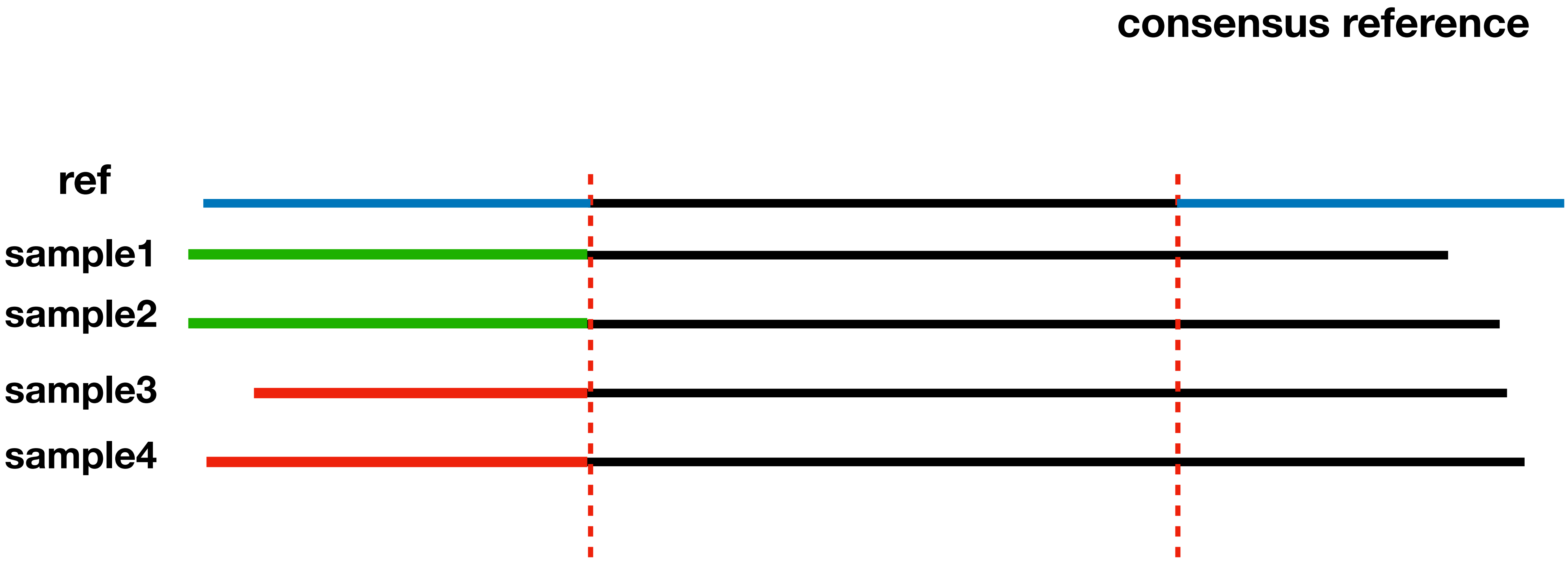
## Remove too variable sequences



Remove sequences distant from consensus reference

# Filter poorly aligned regions

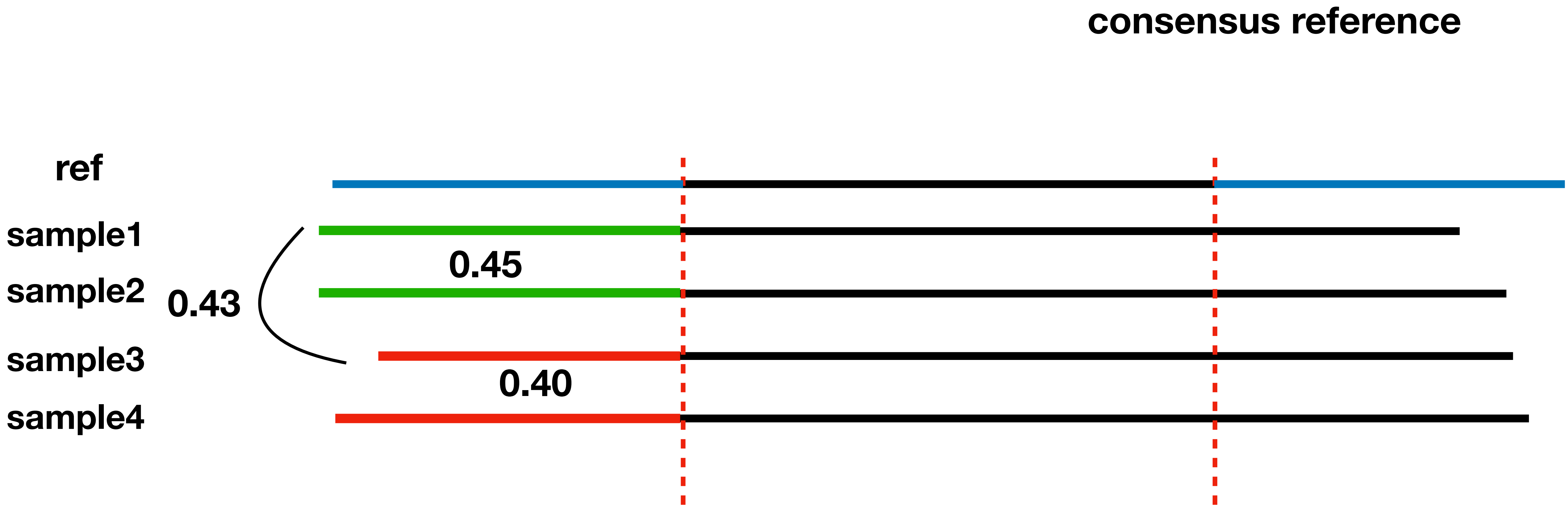
## Remove too variable sequences



2 or more patterns of sequences existing in flanks

# Filter poorly aligned regions

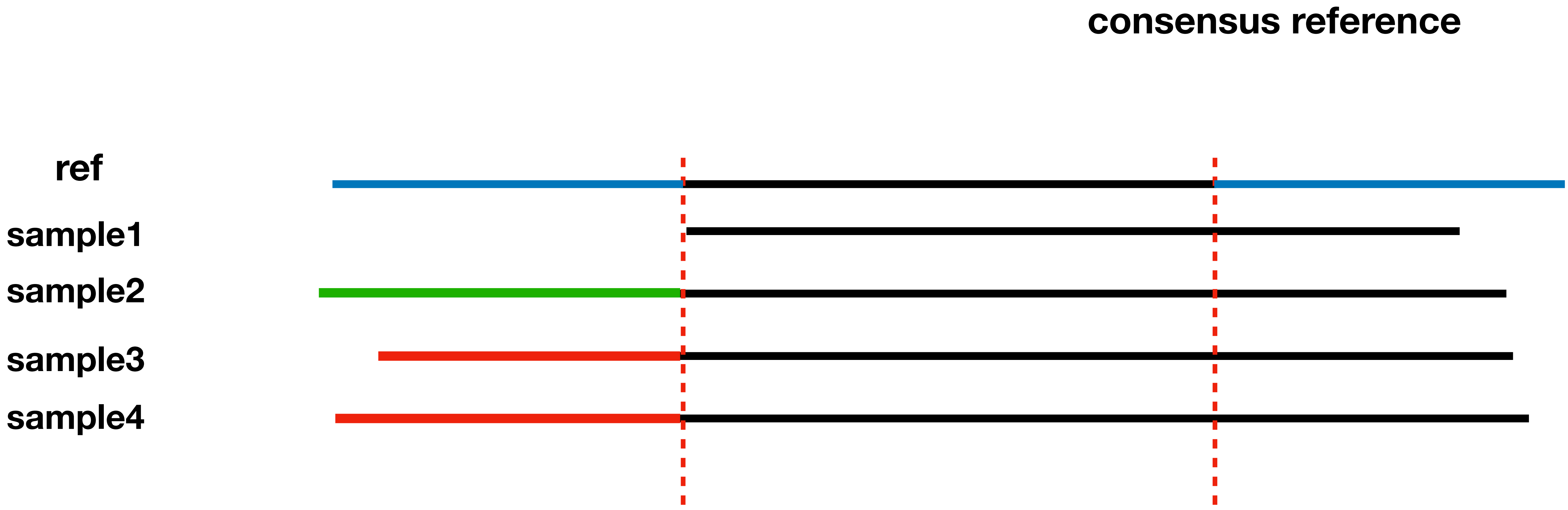
## Remove too variable sequences



Find pair of distant sequences, then compute their distance to the rest of the sequences

# Filter poorly aligned regions

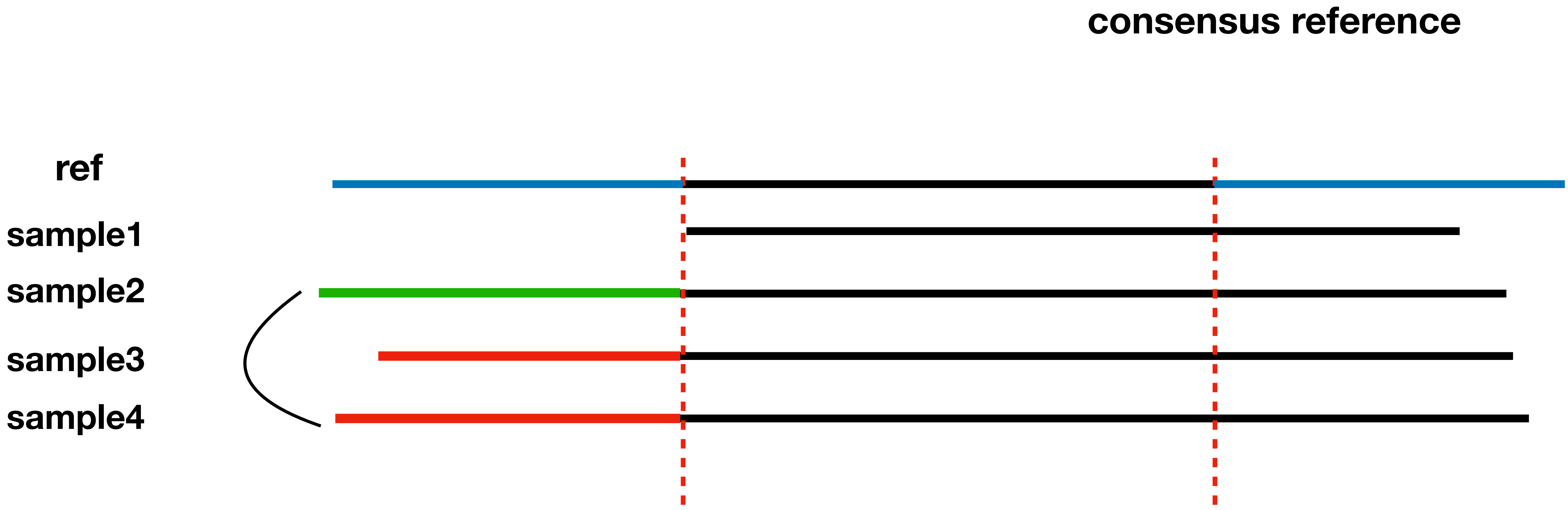
## Remove too variable sequences



Remove sequences distant from rest of the sequences

# Filter poorly aligned regions

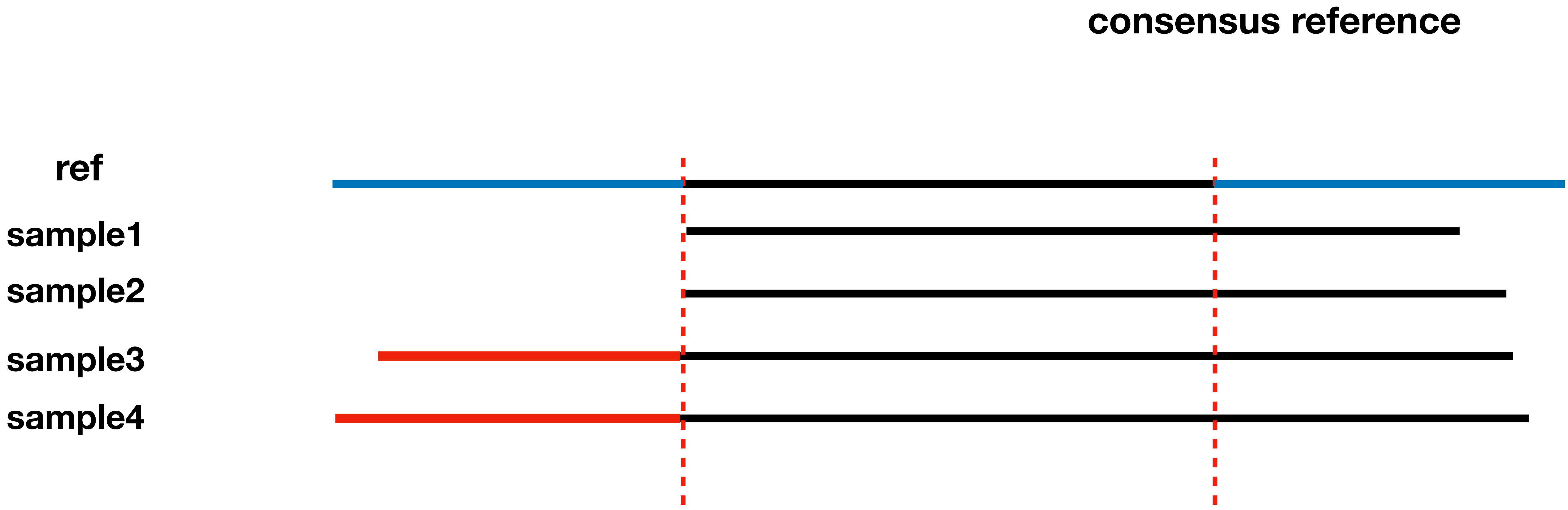
## Remove too variable sequences



Remove sequences distant from rest of the sequences

# Filter poorly aligned regions

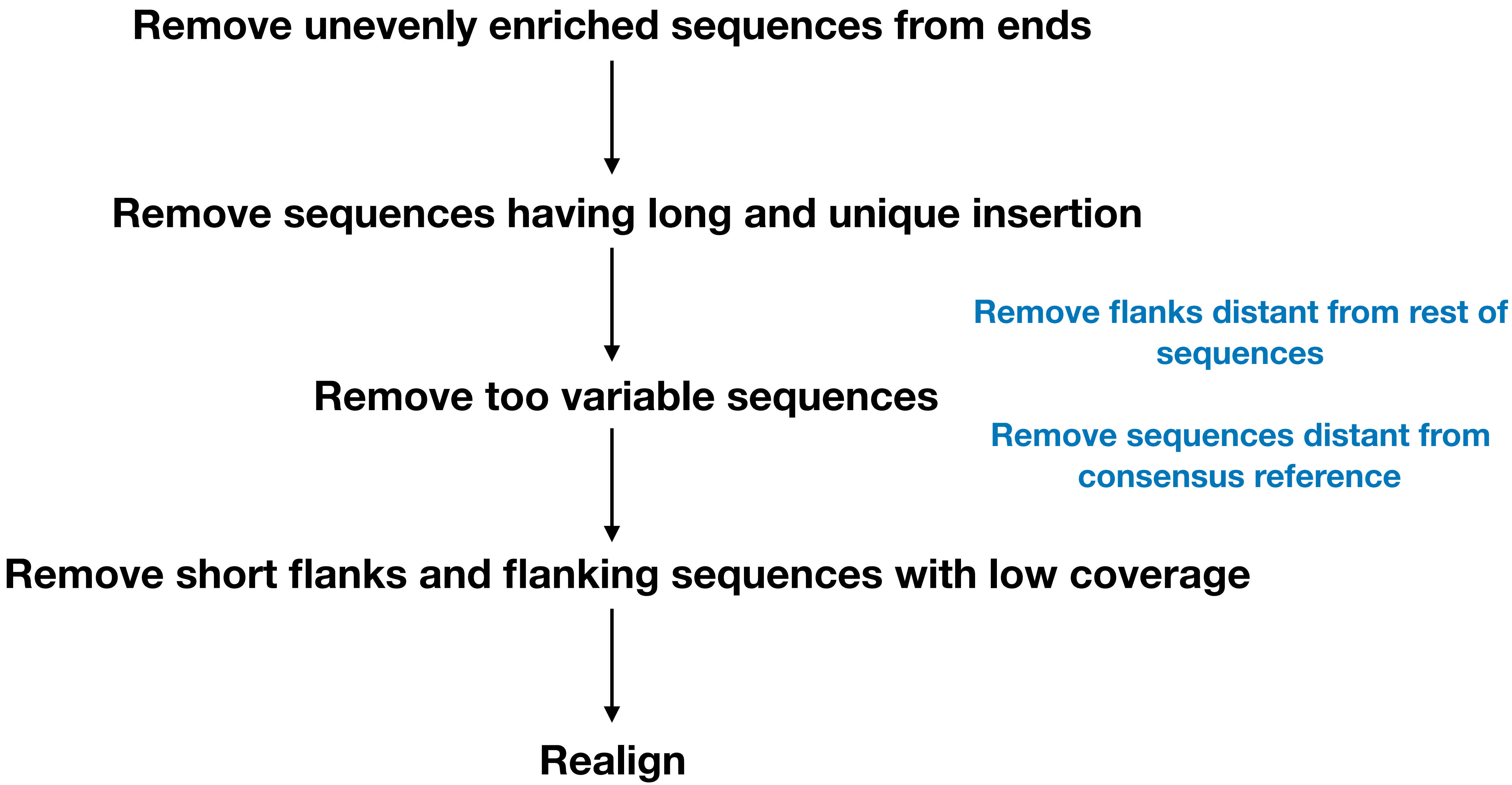
Remove too variable sequences



Remove sequences distant from rest of the sequences

# Filter poorly aligned regions

## Poorly aligned flanking regions



Remove unevenly enriched sequences from ends

Remove sequences having long and unique insertion

Remove too variable sequences

Remove flanks distant from rest of sequences

Remove sequences distant from consensus reference

Remove short flanks and flanking sequences with low coverage

Realign

# Filter poorly aligned regions

## Poorly aligned flanking regions

## Usage:

```
$ perl flank_filter.pl \
--flank assemble_result/f \                          ← sequences with flanks
--nonflank_filtered filtered_nf \                    ← filtered coding sequences
--flank_filtered Oreochromis_niloticus \             ← filtered sequences with flanks
--ref_taxa Oreochromis_niloticus                     ← name of reference
```

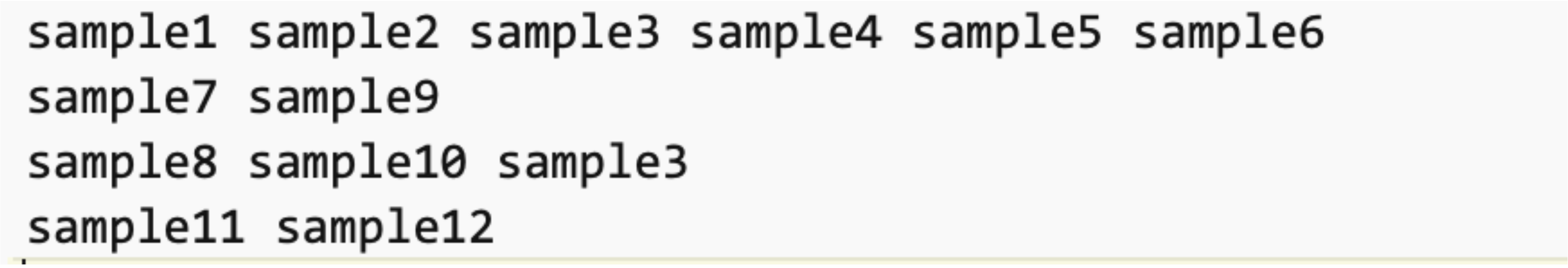
# Filter for other purpose

Filter out loci with pre-defined monophyletic group

Filter out loci follow the molecular clock hypotheses

# Filter for other purpose

## Filter out loci with pre-defined monophyletic group



```
sample1 sample2 sample3 sample4 sample5 sample6
sample7 sample9
sample8 sample10 sample3
sample11 sample12
```

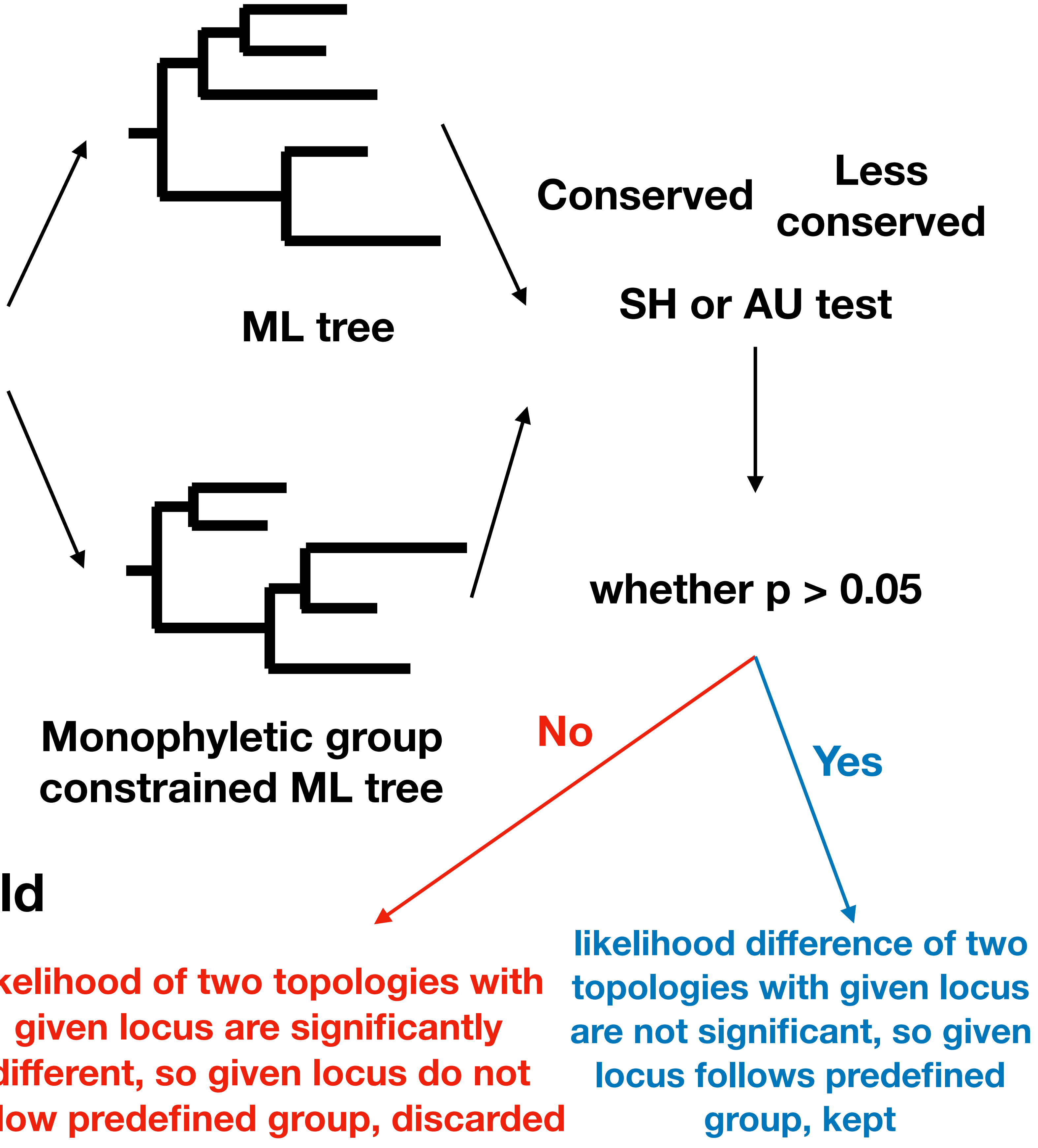
**Define monophyletic group in a txt file, one group a line**

**At least two sample in a group**

# Filter for other purpose

## SH and AU test

```
ATCGTAGGGCTGGCTAGTCGTAGCTA
ATCGTAGGGCTGGCGAGTCGT-GCTA
ATCGTAGGACTGGCTAGTCGTAGCTA
ATCGTACGGCTGGCTAGTCGTAGCTA
ATCGTAGGGCTGGCTAGTCGTAGCTA
```



ML tree

Monophyletic group constrained ML tree

Conserved

Less conserved

SH or AU test

whether p > 0.05

No

Yes

likelihood of two topologies with given locus are significantly different, so given locus do not follow predefined group, discarded

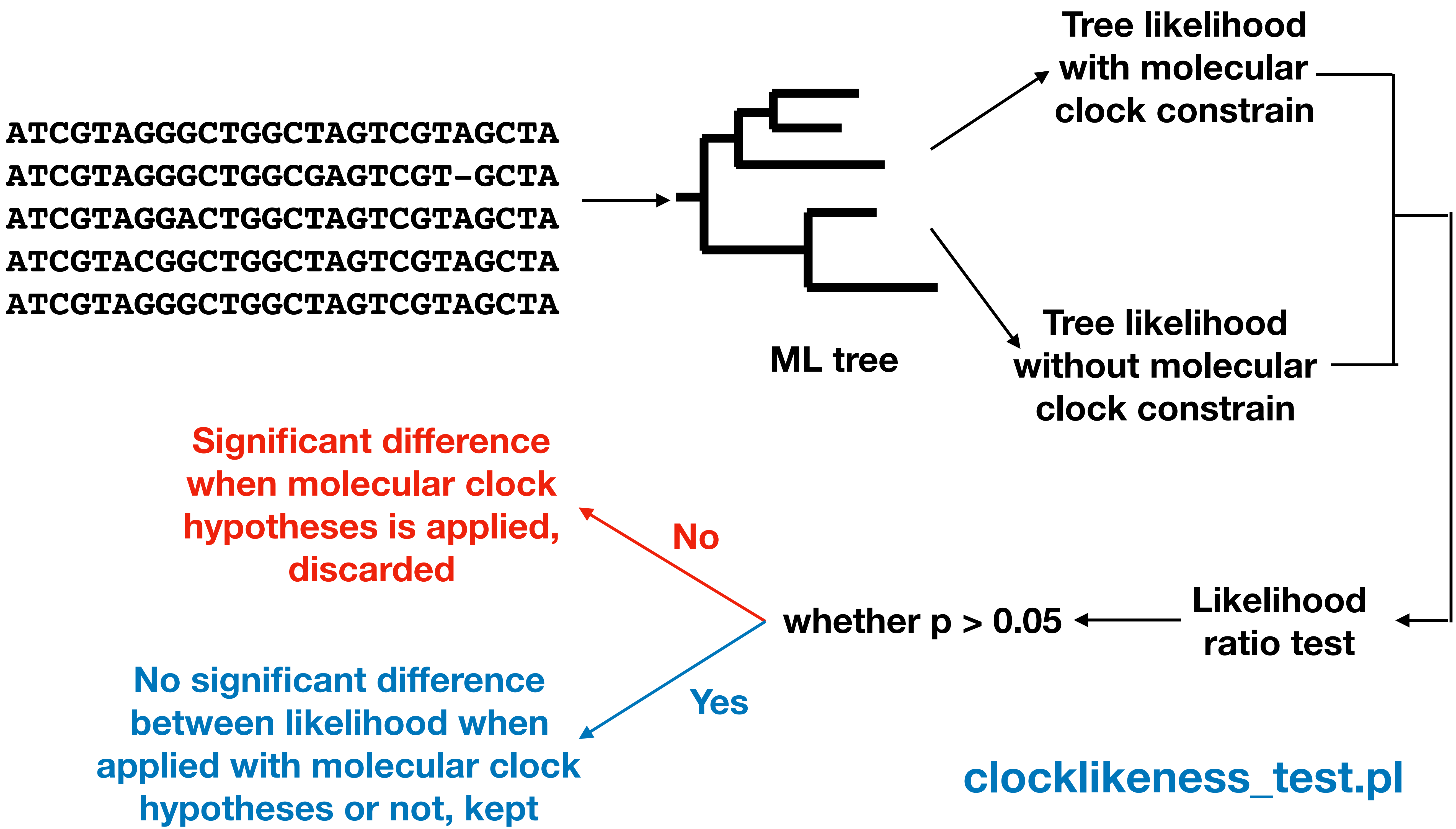
likelihood difference of two topologies with given locus are not significant, so given locus follows predefined group, kept

use construct_tree.pl to build constrained ML tree first

monophyly_test.pl

# Filter for other purpose

## Filter out loci follow the molecular clock hypotheses



ATCGTAGGGCTGGCTAGTCGTAGCTA
ATCGTAGGGCTGGCGAGTCGT-GCTA
ATCGTAGGACTGGCTAGTCGTAGCTA
ATCGTACGGCTGGCTAGTCGTAGCTA
ATCGTAGGGCTGGCTAGTCGTAGCTA

ML tree

Tree likelihood with molecular clock constrain

Tree likelihood without molecular clock constrain

Likelihood ratio test

whether p > 0.05

No

Significant difference when molecular clock hypotheses is applied, discarded

Yes

No significant difference between likelihood when applied with molecular clock hypotheses or not, kept

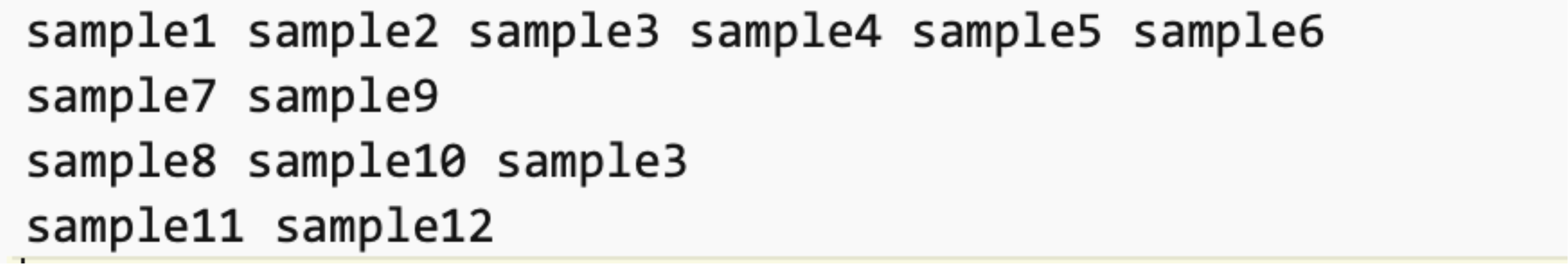
clocklikeness_test.pl

# Detect cross-sample contamination

**Genetic distance of diverged taxa in most loci cannot be very close**

detect_contamination.pl

# Detect cross-sample contamination

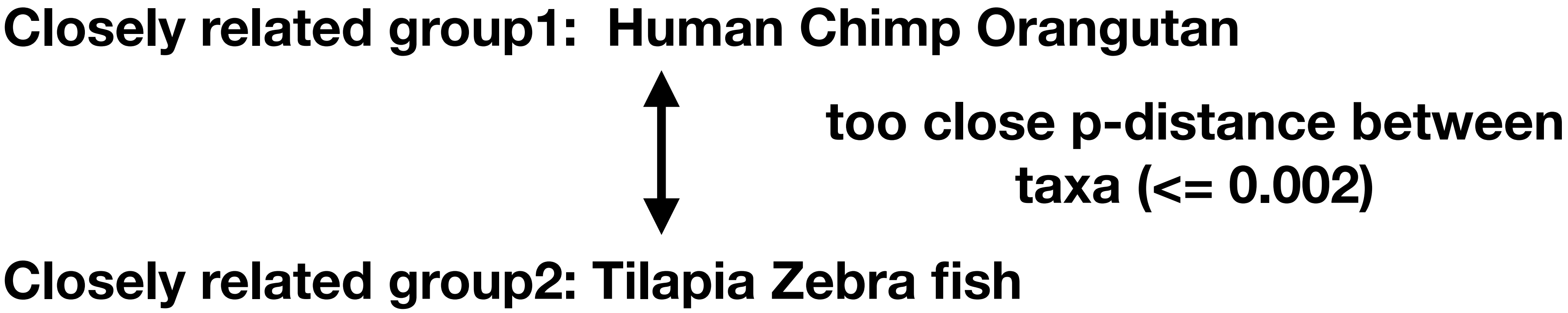


```
sample1 sample2 sample3 sample4 sample5 sample6
sample7 sample9
sample8 sample10 sample3
sample11 sample12
```

**Define close related group in a txt file, one group a line**

**Permit one sample in a group**

# Detect cross-sample contamination



Closely related group1: Human Chimp Orangutan

too close p-distance between taxa (<= 0.002)

Closely related group2: Tilapia Zebra fish

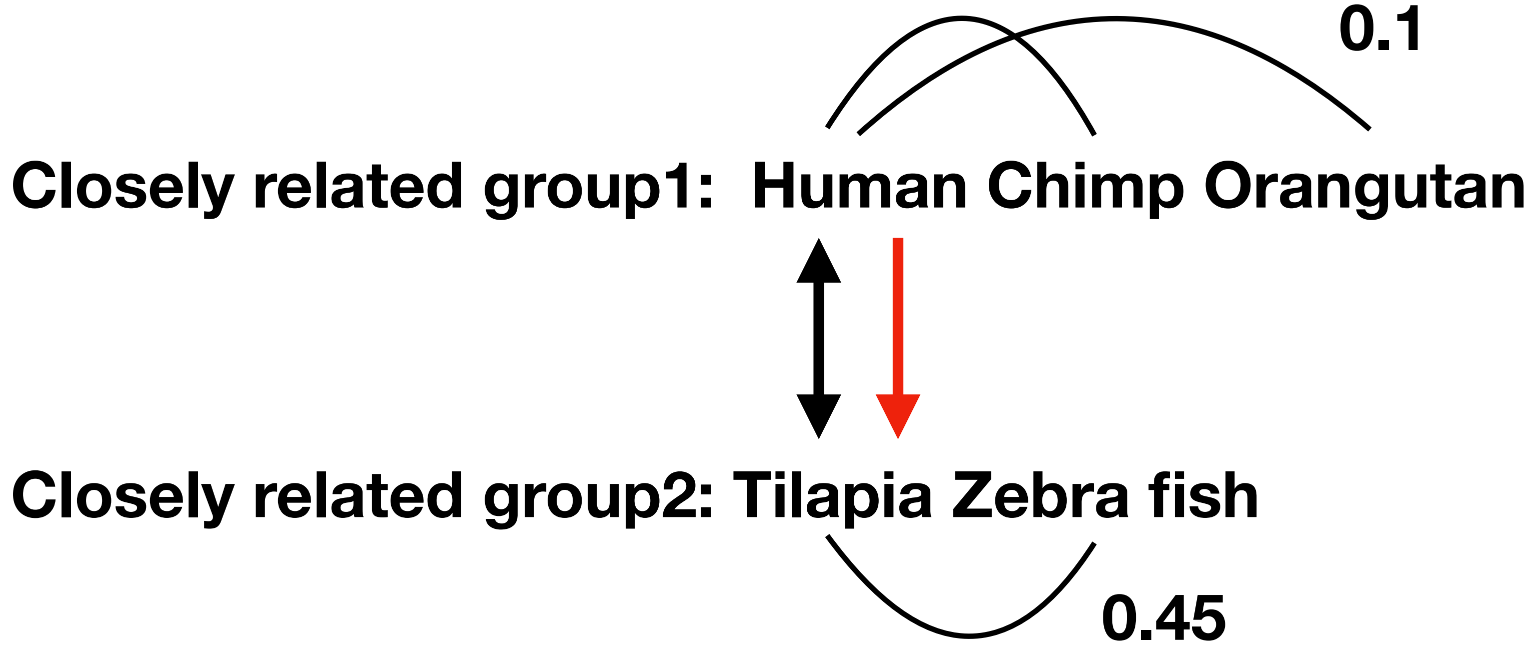
Contamination rate (%) =
Potentially contaminated pair/Co-existence of this pair appeared in all loci*100

4434 loci in total
Potentially contamination between Human and Tilapia among all loci: 10 times
Co-existence of Human-Tilapia pair among all loci: 2000 times

Contamination rate of Human-Tilapia (%) = 10/2000*100 = 0.5%

Default threshold:
>=50% Contamination rate
>= 100 appearing time among all loci

# Detect cross-sample contamination



0.1

Closely related group1: Human Chimp Orangutan

Closely related group2: Tilapia Zebra fish

0.45

Human <- Tilapia: 10
Human -> Tilapia: 200

Sample of Human contaminated Tilapia

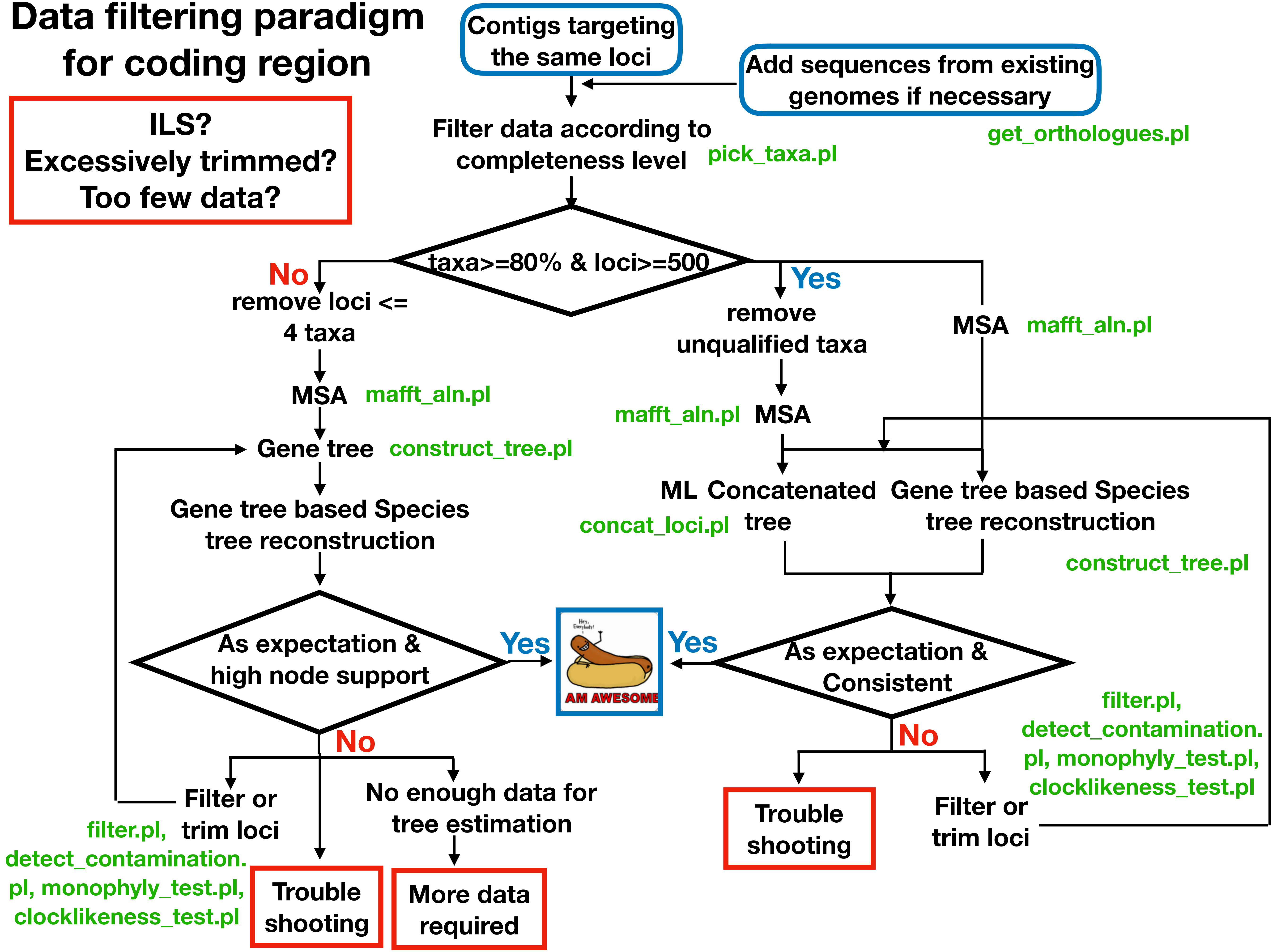
# Summary statistics

Summarized statistics for each locus including:
(1) Average length of coding region
(2) Average length of flanking region
(3) Length of alignment
(4) Average GC content
(5) Percentage of Missing data
(6) Pairwise distance

Summarized statistics for each sample including:
(1) Average length of captured sequences
(2) Average GC content
(3) Number of captured loci

statistics.pl

# Data filtering paradigm for coding region



ILS?
Excessively trimmed?
Too few data?

Contigs targeting the same loci

Add sequences from existing genomes if necessary

get_orthologues.pl

Filter data according to completeness level

pick_taxa.pl

taxa>=80% & loci>=500

No

remove loci <= 4 taxa

MSA

mafft_aln.pl

Gene tree

construct_tree.pl

Gene tree based Species tree reconstruction

As expectation & high node support

Yes

No

Filter or trim loci

filter.pl, detect_contamination.pl, monophyly_test.pl, clocklikeness_test.pl

Trouble shooting

No enough data for tree estimation

More data required

Yes

remove unqualified taxa

mafft_aln.pl

MSA

MSA

mafft_aln.pl

ML Concatenated tree

concat_loci.pl

Gene tree based Species tree reconstruction

construct_tree.pl

As expectation & Consistent

Yes

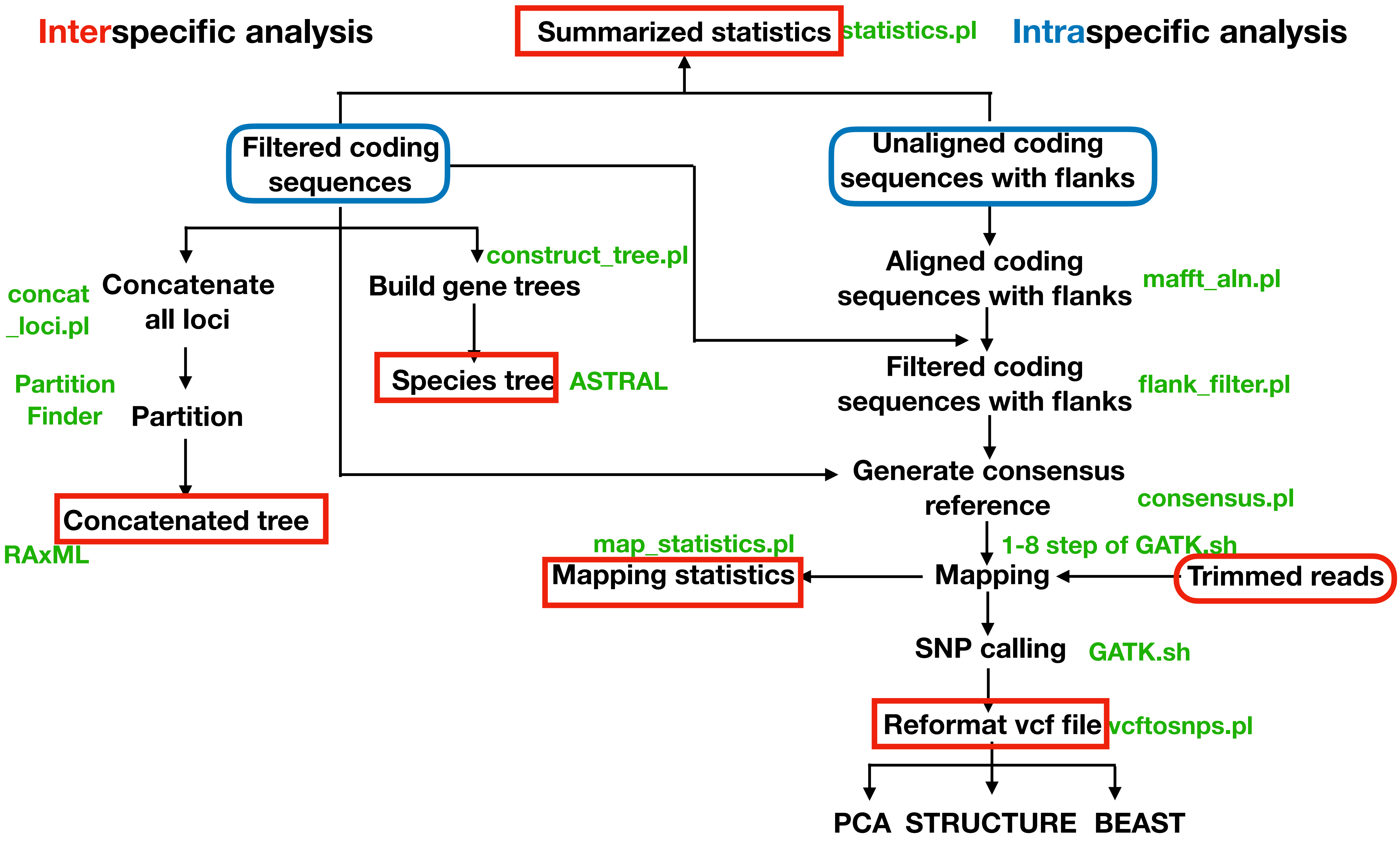
No

Trouble shooting

Filter or trim loci

filter.pl, detect_contamination.pl, monophyly_test.pl, clocklikeness_test.pl

# Conventional analysis paradigm


Interspecific analysis
Summarized statistics
statistics.pl
Intraspecific analysis
Filtered coding sequences
Unaligned coding sequences with flanks
concat_loci.pl
Concatenate all loci
construct_tree.pl
Build gene trees
Aligned coding sequences with flanks
mafft_aln.pl
Partition Finder
Partition
Species tree
ASTRAL
Filtered coding sequences with flanks
flank_filter.pl
Generate consensus reference
consensus.pl
Concatenated tree
RAxML
map_statistics.pl
Mapping statistics
1-8 step of GATK.sh
Mapping
Trimmed reads
SNP calling
GATK.sh
Reformat vcf file
vcftosnps.pl
PCA
STRUCTURE
BEAST